Oak Ridge National Laboratory

Oak Ridge Leadership Computing Facility
Technology Integration Group

Presented by
Sarp Oral, PhD

March 11, 2011

Who are we?

- Bridge builders
  - Act as a bridge between research and production
- Not a directly user-facing group
  - Interacts mostly with HPC Operations, User Assistance and Outreach, and research groups within CSMD
- Supply our expertise directly to projects and other groups in the center
- May be called in to work user problems via Scientific Computing liaisons

Our mission

- Address issues in the OLCF computational environment
  - Bridge the gap between what users need and what is technologically available by third party vendors
  - Bridge the gap between longer-term research and operations
- Plan for future computational platforms and requirements
- Technology evaluations
- Collaboration with vendors on product roadmaps and provide feedback
- Integration of new technologies into OLCF environment

Current Projects and Activities

- Parallel file systems and I/O development
- HPSS development
- Earth System Grid (ESG)
- Common Communication Interface (CCI)
- Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT)

Spider – Persistent storage at the Petascale

Fastest Lustre file system in the world

Demonstrated bandwidth of 240 GB/s on the center-wide file system

Largest scale Lustre file system in the world

Demonstrated stability and concurrent mounts on all major OLCF systems

- Jaguar XT5
- Jaguar XT4
- Frost
- Opteron Dev Cluster (Smokey)
- Visualization Cluster (Lens)

Over 26,000 clients

Hundreds of millions of files

Multiple petabytes of data stored

System	Size	Throughput	OSTs
widow0	4.6 PB	120 GB/s	672
widow1	2.3 PB	60 GB/s	336
widow2	2.3 PB	60 GB/s	336
“sliver”*	1 PB	240 GB/s	1344

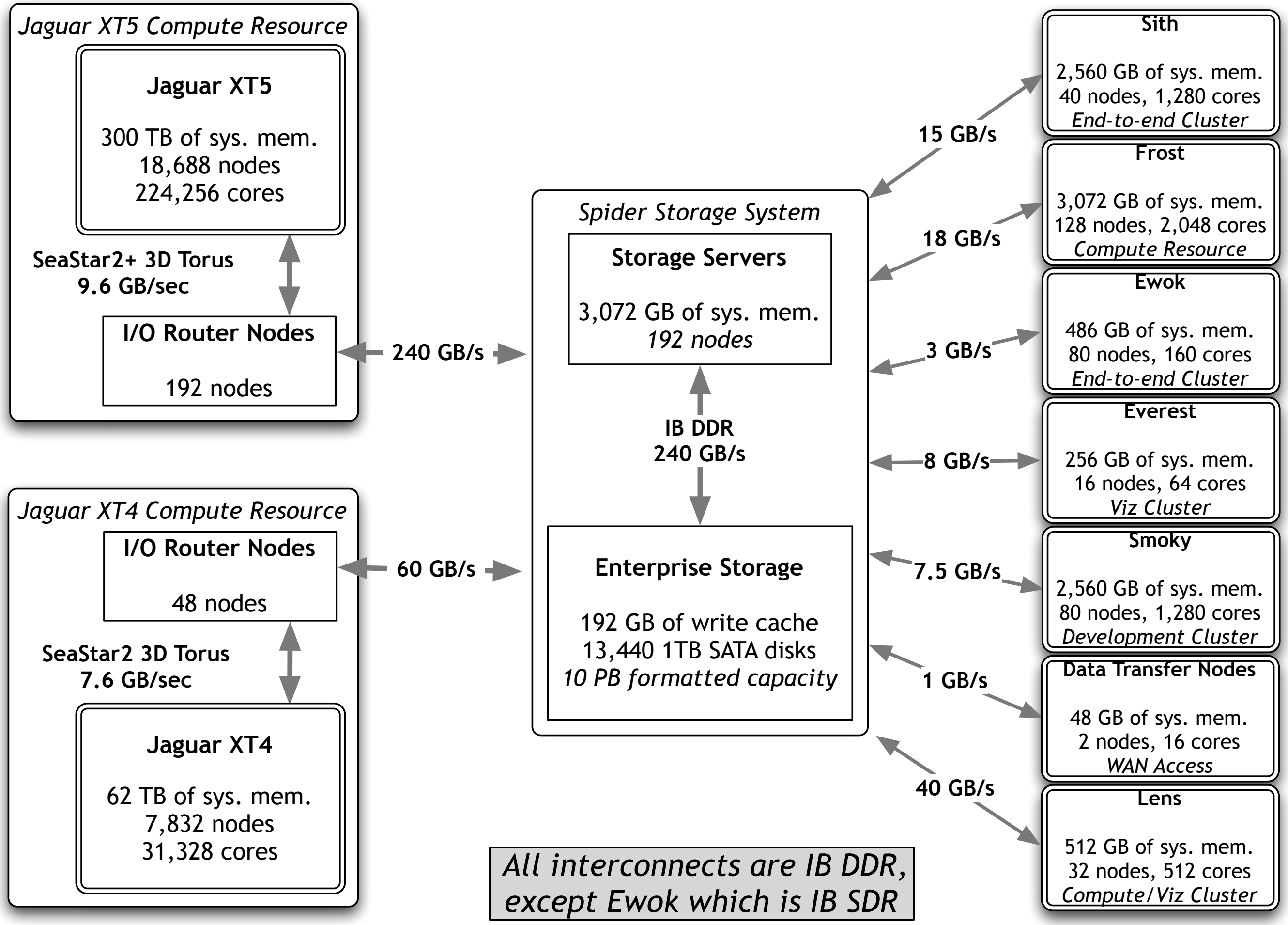
*Not in production

Cutting edge resiliency at scale

Demonstrated resiliency features on Jaguar XT5

- DM Multipath
- Lustre Router failover

Spider – Persistent storage at the Petascale



Jaguar XT5 Compute Resource

Jaguar XT5

300 TB of sys. mem.
18,688 nodes
224,256 cores

SeaStar2+ 3D Torus
9.6 GB/sec

I/O Router Nodes

192 nodes

240 GB/s

Jaguar XT4 Compute Resource

I/O Router Nodes

48 nodes

SeaStar2 3D Torus
7.6 GB/sec

Jaguar XT4

62 TB of sys. mem.
7,832 nodes
31,328 cores

60 GB/s

Spider Storage System

Storage Servers

3,072 GB of sys. mem.
192 nodes

IB DDR
240 GB/s

Enterprise Storage

192 GB of write cache
13,440 1TB SATA disks
10 PB formatted capacity

15 GB/s

Sith

2,560 GB of sys. mem.
40 nodes, 1,280 cores
End-to-end Cluster

18 GB/s

Frost

3,072 GB of sys. mem.
128 nodes, 2,048 cores
Compute Resource

3 GB/s

Ewok

486 GB of sys. mem.
80 nodes, 160 cores
End-to-end Cluster

8 GB/s

Everest

256 GB of sys. mem.
16 nodes, 64 cores
Viz Cluster

7.5 GB/s

Smoky

2,560 GB of sys. mem.
80 nodes, 1,280 cores
Development Cluster

1 GB/s

Data Transfer Nodes

48 GB of sys. mem.
2 nodes, 16 cores
WAN Access

40 GB/s

Lens

512 GB of sys. mem.
32 nodes, 512 cores
Compute/Viz Cluster

All interconnects are IB DDR, except Ewok which is IB SDR

Parallel file systems and I/O development

- Improved Lustre metadata performance and system resiliency
- OpenSFS development
- OLCF-3 testbed evaluation
  - Low-level, bare metal, evaluation of new file and storage technologies for OLCF-3 environment
    - Collaboration with file and storage system vendors
    - New technologies such as Flash, faster RAID array builds

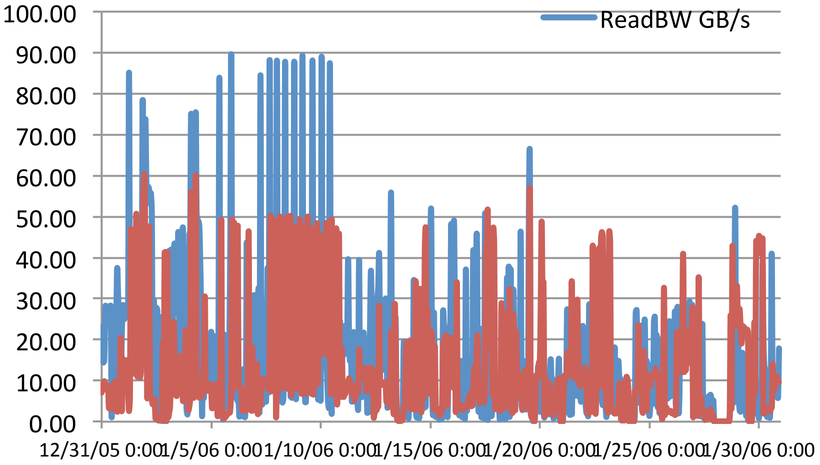
whamcloud

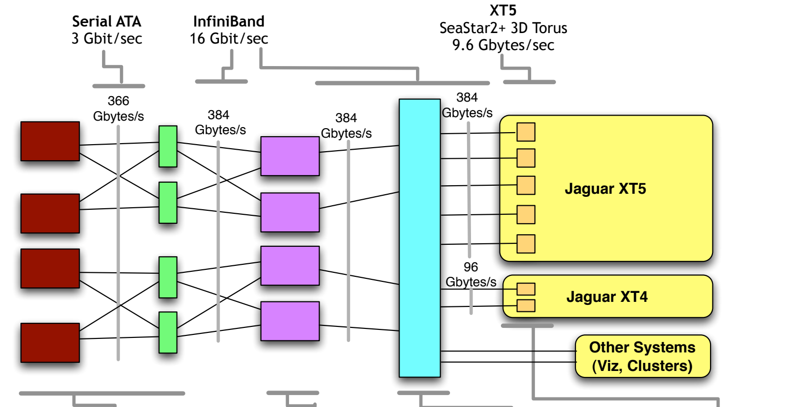
OpenSFS

LSI VIRIDENT TERASCALA DataDirect NETWORKS

Parallel file systems and I/O development

- Lustre Development
  - Feature Enhancements
  - Bug Fixes
- Operational Improvements
  - Metadata Performance
  - Monitoring and Diagnostic Tools
- Parallel Data Tools
  - LSQ (Quicker LS for Lustre)
  - SPDCP (Parallel Copy)
  - PLTAR (Parallel Tar)
  - IOTA (I/O Tracing)





Enterprise Storage
controllers and large racks of disks are connected via InfiniBand.

48 DataDirect S2A9900 controller pairs with 1 Tbyte drives and 4 InifiniBand connections per pair

Storage Nodes
run parallel file system software and manage incoming FS traffic.

192 dual quad core Xeon servers with 16 Gbytes of RAM each

SION Network
provides connectivity between OLCF resources and primarily carries storage traffic.

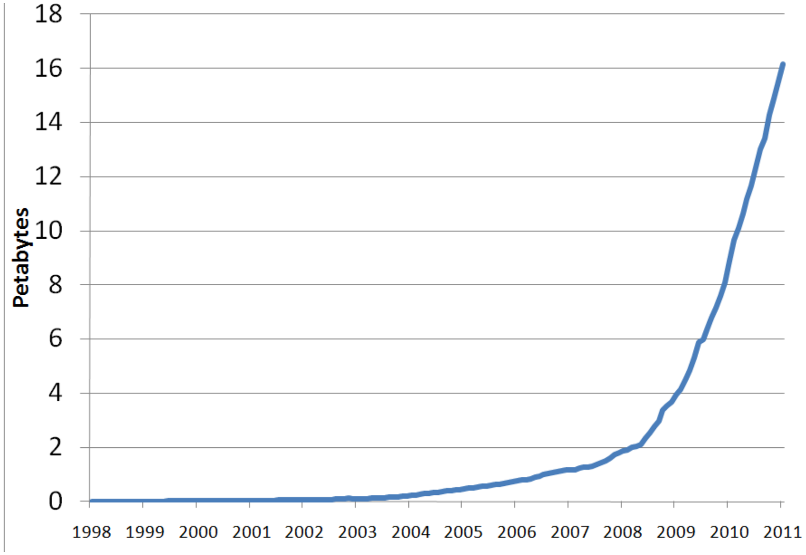
3000+ port 16 Gbit/sec InfiniBand switch complex

Lustre Router Nodes
run parallel file system client software and forward I/O operations from HPC clients.

192 (XT5) and 48 (XT4) one dual core Opteron nodes with 8 GB of RAM each

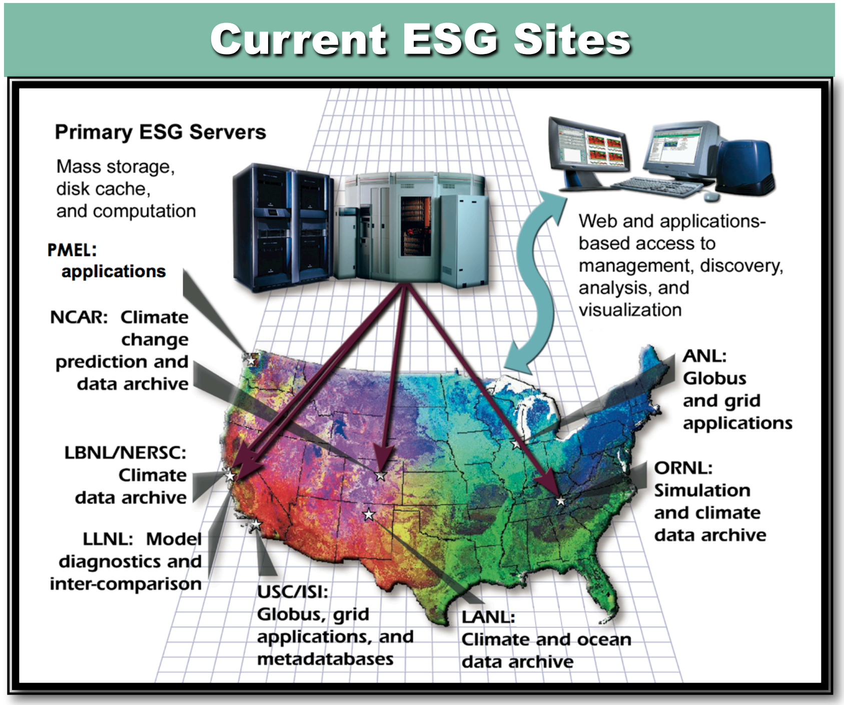
HPSS Development

- Core Contributor to HPSS
  - Storage System Manager
  - Logging Subsystem
  - Bitfile Server
  - Accounting Subsystem
- HPSS operations
  - Production support
- Gearing up For HPSS 8.1
  - Major architectural changes
  - Targeted to meet our requirements for archival storage through 2016
  - > 640 Petabytes
  - > ¼ billion files
  - > 500 GB/sec



Earth Systems Grid (ESG)

- Core competencies in federated data management
  - Developed end-to-end mechanism to publish datasets within NCCS HPSS to the public
  - Support for Observational Datasets (ARM, CDIAC)
  - Next generation portal design and development
- Leveraging the ESG infrastructure
  - Provide data portals for all types of scientific data (beyond climate) within HPSS archives and disk cache



Current ESG Sites

Primary ESG Servers

Mass storage, disk cache, and computation

Web and applications-based access to management, discovery, analysis, and visualization

PMEL: applications

NCAR: Climate change prediction and data archive

LBNL/NERSC: Climate data archive

LLNL: Model diagnostics and inter-comparison

USC/ISI: Globus, grid applications, and metadatabases

LANL: Climate and ocean data archive

ANL: Globus and grid applications

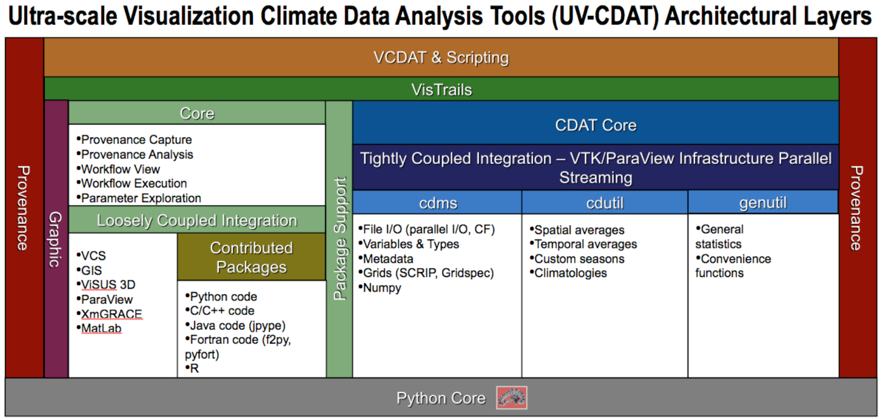
ORNL: Simulation and climate data archive

Common Communication Interface (CCI)

- Developing a common interface for various high-performance networking technologies
  - Cray Portals, Cray Gemini, 10G Ethernet, Infiniband
  - Ability to bridge heterogeneous networks
- Facilitate ease of use without sacrificing performance
  - As easy as using Sockets
  - Performance on-par with low level interfaces
    - (low-latency, zero-copy)
  - Portable across all networking technologies of interest
  - Scalable to leadership class systems
- Will support a wide variety of parallel tools, runtimes, monitoring systems, etc.
- Nearing completion of our prototype implementation

UV-CDAT development

- Ultra-Scale Visualization – Climate Data Analysis Tools
  - Developing state-of-the-art tools to support BER (Office of Biological & Environmental Research) climate research



Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) Architectural Layers

VCDAT & Scripting

VisTrails

Core

- Provenance Capture
- Provenance Analysis
- Workflow View
- Workflow Execution
- Parameter Exploration

Loosely Coupled Integration

- VCS
- GIS
- ViSUS 3D
- ParaView
- XmGRACE
- MatLab

Contributed Packages

- Python code
- C/C++ code
- Java code (jpype)
- Fortran code (f2py, pyfort)
- R

CDAT Core

Tightly Coupled Integration – VTK/ParaView Infrastructure Parallel Streaming

cdms

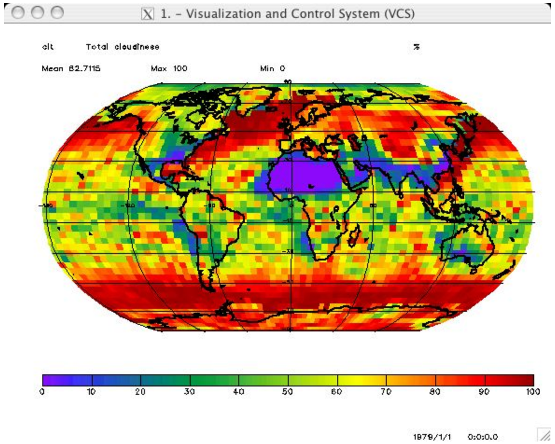
- File I/O (parallel I/O, CF)
- Variables & Types
- Metadata
- Grids (SCRIP, Gridspec)
- Numpy

cdutil

- Spatial averages
- Temporal averages
- Custom seasons
- Climatologies

genutil

- General statistics
- Convenience functions



Questions?

- Contact info:

  Galen M. Shipman

  Group Leader

  865-576-2672

  gshipman@ornl.gov